Relative Information Loss in the PCA

Bernhard C. Geiger*, Gernot Kubin*

*Signal Processing and Speech Communication Laboratory, Graz University of Technology, Austria

{geiger,gernot.kubin}@tugraz.at

Abstract—In this work we analyze principle component analysis (PCA) as a deterministic input-output system. We show that the relative information loss induced by reducing the dimensionality of the data after performing the PCA is the same as in dimensionality reduction without PCA. Furthermore, we analyze the case where the PCA uses the sample covariance matrix to compute the rotation. If the rotation matrix is not available at the output, we show that an infinite amount of information is lost. The relative information loss is shown to decrease with increasing sample size.

I. Introduction

Principle component analysis (PCA) is a powerful tool for both linear decorrelation and dimensionality reduction, and is thus widely used in machine learning, neural networks, and feature extraction [1], [2]. A vast literature proves the optimality of PCA in an information theoretic sense for specific cases: Linsker [3] proved that for N-dimensional Gaussian data corrupted by Gaussian noise with diagonal covariance matrix, the PCA maximizes the mutual information between the data and a one-dimensional output random variable. Plumbley argued in [4] that for non-Gaussian data the PCA is the linear transform which minimizes an upper bound on the information lost due to dimensionality reduction. Furthermore, the authors of [1] show that under some circumstances the PCA minimizes redundancy in the output data, i.e., is an optimal linear independent component analysis. While this list of previous works is clearly not complete, it highlights the utility of information-theoretic measures to characterize the performance of PCA algorithms.

In the present work, we pursue a slightly different approach: We view the PCA as a multivariate, vector-valued input-output system and analyze it in terms of information loss. In this context, *information* means the *total information* available at the input of the system, in contrast to [3], [4], which considered only information which is *relevant*. As a consequence, while our analysis may superficially appear to contradict the results available in the literature, it rather provides a different view on PCA as an information-processing system.

In its standard notation, PCA is a linear system

$$\mathbf{Y} = \underline{\mathbf{W}}^T \mathbf{X} \tag{1}$$

where $\mathbf{X}$ and $\mathbf{Y}$ are the N-dimensional input and output vectors, respectively, and $\underline{\mathbf{W}}$ is an orthogonal matrix. In this sense, the PCA is a bijective transform and thus invertible. Typically, however, the output vector will be truncated after performing the PCA in order to reduce the dimensionality of the data. In this projection to a lower-dimensional subspace

information is lost, and we intend to quantify the loss in this work.

Let us be more specific: Assuming continuous-valued random variables (RV), the information available at the input is infinite. In particular, assuming joint continuity, each component of the multi-dimensional input RV contains an infinite amount of information; thus, if the dimension is reduced, i.e., if some components are dropped, an infinite amount of information is lost (cf. [5]).

In case the orthogonal matrix is not known a priori but has to be estimated from a set of input data vectors collected in the matrix $\underline{\mathbf{X}}$, the PCA becomes a nonlinear operation:

$$\underline{\mathbf{Y}} = \underline{\mathbf{w}}(\underline{\mathbf{X}})\underline{\mathbf{X}}. \tag{2}$$

Here, $\underline{\mathbf{w}}$ is a matrix-valued function which computes the orthogonal matrix required for rotating the data (e.g., using the QR algorithm [6]). If this orthogonal matrix is not stored and made available at the output, one will readily agree that information is lost even if – at a first glance – the dimension of the data is not reduced.

Our notion of *relative* information loss captures the ratio of input information which cannot be retrieved at the output. We make this statement precise in Section II. Before analyzing the information loss in PCA in Section V, we present a general theorem about the loss in systems which reduce the dimensionality of the data (Section III). Section IV acts as a bridge, containing some toy examples which should give an intuitive understanding of relative information loss. In Section VI, the PCA using an input data matrix (cf. (2)) to perform the rotation is shown to destroy information even if the dimensionality is not reduced. Since none of our findings explains the usefulness of the PCA, we finally give an outlook of how to corroborate its optimality using a different notion of information loss in Section VII. We there also discuss possible implications for a system theory from an information-theoretic point-of-view.

II. Relative Information Loss – A Quick Overview

In this section we will briefly present the basic properties of relative information loss. To this end, we introduce

Definition 1 (Relative Information Loss)**.** Let $\mathbf{X}$ be an N-dimensional RV valued in $\mathcal{X}$, and let $\mathbf{Y}$ be obtained by transforming $\mathbf{X}$ with a static function g, i.e., $\mathbf{Y} = g(\mathbf{X})$. We define the *relative information loss* induced by this transform as

$$l(\mathbf{X} \to \mathbf{Y}) = \lim_{n \to \infty} \frac{H(\hat{\mathbf{X}}_n | \mathbf{Y})}{H(\hat{\mathbf{X}}_n)} \tag{3}$$

where $\hat{\mathbf{X}}_n = \frac{\lfloor n\mathbf{X} \rfloor}{n}$ (element-wise). The quantity on the left is defined if the limit on the right-hand side exists.

Loosely speaking, this quantity represents the percentage of input information which is lost in the system, with every bit weighted equally. Taking, for example, an 8 bit number[1], losing the most significant or the least significant bit both amounts to a relative information loss of $l(X \to Y) = \frac{1}{8}$. The advantage of this definition is its independence of application-specific aspects (where, e.g., the most significant bit may be more important than the least significant one; cf. Section VII for a discussion).

The motivation for introducing this quantity is to complement the absolute notion of information loss, given by $L(\mathbf{X} \to \mathbf{Y}) = H(\mathbf{X}|\mathbf{Y})$, which suffers from being infinite in many practically relevant cases (cf. [5]).

Due to the non-negativity of entropy and the fact that conditioning reduces entropy, it follows that $l(\mathbf{X} \to \mathbf{Y}) \in [0, 1]$. Moreover, for continuous input RVs $\mathbf{X}$, a nonzero relative loss corresponds to an infinite absolute loss:

Proposition 1. *Let $\mathbf{X}$ be such that $H(\mathbf{X}) = \infty$ and let $l(\mathbf{X} \to \mathbf{Y}) > 0$. Then, $L(\mathbf{X} \to \mathbf{Y}) = H(\mathbf{X}|\mathbf{Y}) = \infty$.*

Proof: We prove the proposition by contradiction. To this end, assume that $H(\mathbf{X}|\mathbf{Y}) = L \leq \infty$. Then,

$$l(\mathbf{X} \to \mathbf{Y}) = \lim_{n \to \infty} \frac{H(\hat{\mathbf{X}}_n|\mathbf{Y})}{H(\hat{\mathbf{X}}_n)} = \lim_{n \to \infty} \inf \frac{H(\hat{\mathbf{X}}_n|\mathbf{Y})}{H(\hat{\mathbf{X}}_n)} \quad (4)$$

$$\leq \lim_{n \to \infty} \inf \frac{H(\mathbf{X}|\mathbf{Y})}{H(\hat{\mathbf{X}}_n)} \quad (5)$$

$$= \lim_{n \to \infty} \inf \frac{L}{H(\hat{\mathbf{X}}_n)} = 0 \quad (6)$$

where the inequality is due to data processing. The last equality follows from the fact that at least a subsequence of $H(\hat{\mathbf{X}}_n)$ converges to $H(\mathbf{X})$ (cf. [7]). ∎

Furthermore, we can show a tight connection to the information dimension introduced by Rényi [8]. We therefore need the following

Definition 2 (Information Dimension [8])**.** The *information dimension* of an RV $\mathbf{X}$ is given as

$$d(\mathbf{X}) = \lim_{n \to \infty} \frac{H(\hat{\mathbf{X}}_n)}{\log n} \quad (7)$$

provided the limit on the right exists.

Rényi then showed that the asymptotic behavior of $H(\hat{\mathbf{X}}_n)$ depends strongly on the information dimension of $\mathbf{X}$:

Lemma 1 (Asymptotic behavior of $H(\hat{\mathbf{X}}_n)$)**.** *Let $\mathbf{X}$ be an RV with existing information dimension $d(\mathbf{X})$ and let $H(\hat{\mathbf{X}}_1) < \infty$. Then, for $n \to \infty$ the entropy of $\hat{\mathbf{X}}_n$ behaves as*

$$H(\hat{\mathbf{X}}_n) = d(\mathbf{X}) \log n + h + o(1) \quad (8)$$

---

[1]I.e., X is a scalar RV which can be represented by eight independent Bernoulli-$\frac{1}{2}$ RVs; thus $H(X) = 8$ bit.

where h is the $d(\mathbf{X})$-dimensional entropy of $\mathbf{X}$ (provided it exists).

Proof: See [8] (cf. also [9], [10]). ∎

In particular, if $\mathbf{X}$ has a probability measure absolutely continuous w.r.t. the N-dimensional Lebesgue measure μ^N ($P_\mathbf{X} \ll \mu^N$), then the information dimension $d(\mathbf{X}) = N$ and h denotes the differential entropy of $\mathbf{X}$ (see Theorems 1 & 4 in [8]). We are now ready to make the connection between relative information loss and information dimension in

Theorem 1. *Let $\mathbf{X}$ be an RV with positive (finite) information dimension $d(\mathbf{X})$. Then, if $d(\mathbf{X}|\mathbf{Y} = \mathbf{y})$ exists and is finite $P_\mathbf{Y}$-a.s., the relative information loss equals*

$$l(\mathbf{X} \to \mathbf{Y}) = \frac{d(\mathbf{X}|\mathbf{Y})}{d(\mathbf{X})} \quad (9)$$

where $d(\mathbf{X}|\mathbf{Y}) = \int_\mathcal{Y} d(\mathbf{X}|\mathbf{Y} = \mathbf{y}) dP_\mathbf{Y}(\mathbf{y})$.

Proof: We start with Definition 1 and obtain

$$l(\mathbf{X} \to \mathbf{Y}) = \lim_{n \to \infty} \frac{H(\hat{\mathbf{X}}_n|\mathbf{Y})}{H(\hat{\mathbf{X}}_n)} \quad (10)$$

$$= \lim_{n \to \infty} \frac{\int_\mathcal{Y} H(\hat{\mathbf{X}}_n|\mathbf{Y} = \mathbf{y}) dP_\mathbf{Y}}{H(\hat{\mathbf{X}}_n)} \quad (11)$$

$$= \lim_{n \to \infty} \frac{\int_\mathcal{Y} \frac{H(\hat{\mathbf{X}}_n|\mathbf{Y}=\mathbf{y})}{\log n} dP_\mathbf{Y}}{\frac{H(\hat{\mathbf{X}}_n)}{\log n}} \quad (12)$$

where we divided both the numerator and the denominator by $\log n$. Assuming now the limits of the numerator and denominator exist and are finite we can continue with

$$l(\mathbf{X} \to \mathbf{Y}) = \frac{\lim_{n \to \infty} \int_\mathcal{Y} \frac{H(\hat{\mathbf{X}}_n|\mathbf{Y}=\mathbf{y})}{\log n} dP_\mathbf{Y}}{\lim_{n \to \infty} \frac{H(\hat{\mathbf{X}}_n)}{\log n}} \quad (13)$$

$$= \frac{\lim_{n \to \infty} \int_\mathcal{Y} \frac{H(\hat{\mathbf{X}}_n|\mathbf{Y}=\mathbf{y})}{\log n} dP_\mathbf{Y}}{d(\mathbf{X})} \quad (14)$$

where we employed Definition 2. By assumption the limit

$$d(\mathbf{X}|\mathbf{Y} = \mathbf{y}) = \lim_{n \to \infty} \frac{H(\hat{\mathbf{X}}_n|\mathbf{Y} = \mathbf{y})}{\log n} \quad (15)$$

exists $P_\mathbf{Y}$-a.s. Since the information dimension of an RV is upper bounded by the topological dimension of its support (e.g., by the number of coordinates of the vector $\mathbf{X}$), one can apply Lebesgue's dominated convergence theorem (e.g., [11]) to exchange the order of the limit and the integral. This completes the proof. ∎

Before proceeding, we want to mention that the converse of Proposition 1 is not true[2], i.e., that if $H(\mathbf{X}) = \infty$ and $l(\mathbf{X} \to \mathbf{Y}) = 0$, it does not necessarily follow that $L(\mathbf{X} \to \mathbf{Y}) < \infty$. To this end, assume that the function $\mathbf{g}$ is such that, for all $\mathbf{y}$, $\mathbf{X}|\mathbf{Y} = \mathbf{y}$ is a discrete RV with infinite entropy but with $H(\hat{\mathbf{X}}_1|\mathbf{Y} = \mathbf{y}) < \infty$. As a consequence, the

---

[2]We thank an anonymous reviewer for pointing us to this fact.

information dimension $d(\mathbf{X}|\mathbf{Y} = \mathbf{y}) = 0$, which establishes $l(\mathbf{X} \to \mathbf{Y}) = 0$. However,

$$L(\mathbf{X} \to \mathbf{Y}) = \int_{\mathcal{Y}} H(\mathbf{X}|\mathbf{Y} = \mathbf{y})dP_{\mathbf{Y}}(\mathbf{y}) = \infty. \quad (16)$$

We give an example for such a case in the Appendix.

III. RELATIVE INFORMATION LOSS FOR FUNCTIONS WHICH REDUCE DIMENSIONALITY

We now proceed to analyzing the relative information loss for measurable functions $\mathbf{g}\colon \mathcal{X} \to \mathcal{Y}$ which map subsets of $\mathcal{X} \subseteq \mathbb{R}^N$ to sufficiently well-behaved submanifolds of $\mathbb{R}^N$. In particular, let $\{\mathcal{X}_i\}$, $i = 1, \dots, L$, denote a finite partition of $\mathcal{X}$ and let $P_{\mathbf{X}} \ll \mu^N$, where μ^N is the N-dimensional Lebesgue measure. Here and throughout the remainder of this paper we assume that all involved information dimensions exist and are finite. The latter restriction is fulfilled under the mild condition that $H(\hat{\mathbf{X}}_n) < \infty$ [8], which for scalar (one-dimensional) RVs X translates to $\mathrm{E}\{|X|^\epsilon\} < \infty$ for some $\epsilon > 0$ [12].

We maintain

Theorem 2. *Let $\mathbf{X}$ be such that $P_{\mathbf{X}} \ll \mu^N$ is supported on $\mathcal{X} \subseteq \mathbb{R}^N$ and let $\{\mathcal{X}_i\}$ be a partition of $\mathcal{X}$ such that each of its L elements is a smooth N-dimensional manifold. Let $\mathbf{g}$ be such that $\mathbf{g}_i = \mathbf{g}|_{\mathcal{X}_i}$ are submersions to disjoint smooth n_i-dimensional manifolds $\mathcal{Y}_i$. Then, the relative information loss is*

$$l(\mathbf{X} \to \mathbf{Y}) = \sum_{i=1}^{L} P_{\mathbf{X}}(\mathcal{X}_i)\frac{N - n_i}{N}. \quad (17)$$

Proof: By the submersion theorem (e.g., [13]) and the fact that the images of $\mathcal{X}_i$ are disjoint, the preimage of every point in $\mathcal{Y}_i$ is a closed submanifold of $\mathcal{X}_i$ with codimension n_i. We can thus use Theorem 1 together with $d(\mathbf{X}) = N$ to get

$$l(\mathbf{X} \to \mathbf{Y}) = \frac{1}{N}\sum_{i=1}^{L}\int_{\mathcal{Y}_i} d(\mathbf{X}|\mathbf{Y} = \mathbf{y})dP_{\mathbf{Y}}(\mathbf{y})$$

$$= \frac{1}{N}\sum_{i=1}^{L}\int_{\mathcal{Y}_i}(N - n_i)dP_{\mathbf{Y}}(\mathbf{y}) = \sum_{i=1}^{L}\frac{N - n_i}{N}P_{\mathbf{Y}}(\mathcal{Y}_i) \quad (18)$$

which completes the proof by noticing that $P_{\mathbf{Y}}(\mathcal{Y}_i) = P_{\mathbf{X}}(\mathcal{X}_i)$. ∎

Before proceeding, two interesting facts are worth mentioning: First of all, as a submersion is a smooth mapping between smooth manifolds, we can state

Corollary 1. *If $P_{\mathbf{X}} \ll \mu^N$ and if $\mathbf{g}$ is as in Theorem 2, then*

$$l(\mathbf{X} \to \mathbf{Y}) = 1 - \frac{d(\mathbf{Y})}{d(\mathbf{X})} \quad (19)$$

provided the information dimension of $\mathbf{Y}$ exists.

Proof: We note from the proof of Theorem 2 that

$$l(\mathbf{X} \to \mathbf{Y}) = 1 - \sum_{i=1}^{L}\frac{n_i}{N}P_{\mathbf{Y}}(\mathcal{Y}_i) \quad (20)$$

where $N = d(\mathbf{X})$. We now show that $d(\mathbf{Y}|\mathbf{X} \in \mathcal{X}_i) = n_i$. By the fact that $\mathbf{g}|_{\mathcal{X}_i} \in \mathcal{C}^\infty$ is a submersion, the preimages of μ^{n_i}-null sets are μ^N-null sets themselves [14]. Consequently, since $P_{\mathbf{X}} \ll \mu^N$, the conditional probability measure supported on $\mathcal{Y}_i$ is absolutely continuous w.r.t. the n_i-dimensional Lebesgue measure ($P_{\mathbf{Y}|\mathbf{X} \in \mathcal{X}_i} \ll \mu^{n_i}$). Since $\mathcal{Y}_i$ is a smooth n_i-dimensional manifold, we again obtain with Rényi [8] that $d(\mathbf{Y}|\mathbf{X} \in \mathcal{X}_i) = n_i$. The proof is complete with [12, Theorem 2] or [15], noting that

$$d(\mathbf{Y}) = \sum_{i=1}^{L} d(\mathbf{Y}|\mathbf{X} \in \mathcal{X}_i)P_{\mathbf{X}}(\mathcal{X}_i) = \sum_{i=1}^{L} n_i P_{\mathbf{Y}}(\mathcal{Y}_i). \quad (21)$$

∎

In fact, already Rényi showed (21) for a mixture of a one-dimensional continuous RV and a discrete RV in [8]. For more complicated measures (e.g., measures with a non-integer information dimension) or more general functions it might be hard to prove Theorem 2 and Corollary 1. We conjecture, however, that the conditions imposed in Theorem 2 can be loosened such that at least for a larger class of functions (e.g., those for which the images of $\mathcal{X}_i$ are not disjoint) the relative information loss can be evaluated.

Secondly, we believe that one should be able to analyze cascaded systems, at least in a restricted class of cases: To this end, assume that we have a cascade of two projections: one on the first two and then one on the first coordinate, i.e., $\mathbf{g}\colon [X_1, X_2, X_3] \to [X_1, X_2]$ and $h\colon [X_1, X_2] \to X_1$. Let the RVs X_1, X_2, and X_3 have a continuous joint distribution. Then, according to Theorem 2, the function $\mathbf{g}$ destroys one third of the information, while h destroys half of the remaining information. In total, two thirds of the information are lost. Indeed, we obtain

$$l([X_1, X_2, X_3] \to X_1) = \frac{1}{3} + \frac{1}{2} - \frac{1}{3}\frac{1}{2} = \frac{2}{3}. \quad (22)$$

We are thus lead to the following

Conjecture 1. *Let $\mathbf{g}$ be a (measurable) function describing a system with input $\mathbf{X}$ and output $\mathbf{Y}$. Let further $\mathbf{h}$ be another such system function which transforms input $\mathbf{Y}$ to output $\mathbf{Z}$. There exist conditions under which*

$$l(\mathbf{X} \to \mathbf{Z}) = l(\mathbf{X} \to \mathbf{Y}) + l(\mathbf{Y} \to \mathbf{Z}) - l(\mathbf{X} \to \mathbf{Y})l(\mathbf{Y} \to \mathbf{Z}). \quad (23)$$

Clearly, our example of the cascade of two projections fulfills these conditions. Moreover, as we show in the Appendix, this conjecture holds for discrete RVs with finite entropy.

Finally, it is worth mentioning that the shape of the distribution has no effect on the relative amount of information lost. It is essentially this behavior which leads to the somewhat counter-intuitive results presented in the following sections.

IV. TOY EXAMPLES FOR DIMENSIONALITY REDUCTION

With the help of a few simple examples we now try to make the operational meaning of relative information loss intuitive. At the same time we highlight its importance to the development of an information-centered system theory.
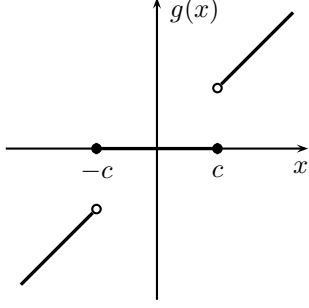
Fig. 1. The center clipper – another example for dimensionality reduction.

Let us introduce a two-dimensional RV $\mathbf{X}$, $P_{\mathbf{X}} \ll \mu^2$ and $d(\mathbf{X}) = 2$, and a transform simply adding the two vector components, i.e.,

$$Y = g(\mathbf{X}) = X_1 + X_2. \qquad (24)$$

We can represent this transform by a cascade of an invertible linear transform (e.g., $\underline{\mathbf{T}} \colon \mathbf{X} \to [X_1 + X_2, X_1]$) and a simple projection onto the first component. Let now $\tilde{\mathbf{X}} = [X_1 + X_2, X_1]$. Since $\underline{\mathbf{T}}$ is invertible and linear, it is bi-Lipschitz and thus preserves the dimension of the transformed RVs. Thus,

$$l(\mathbf{X} \to Y) = l(\tilde{\mathbf{X}} \to Y) = 0.5 \qquad (25)$$

by Theorem 2. Note that this is another example where Conjecture 1 holds.

With this result and the underlying theorems, the theory of information-processing systems can be extended from treating cascade structures as in (23) to also considering parallel structures whose outputs are added. This were not possible by considering only absolute information loss (e.g., [5]), as it would be infinite in these cases.

As a second toy example we consider a center clipper (see Fig. 1), which is commonly used for noise suppression or residual echo cancellation [16]. We describe the center clipper by the following function:

$$g(x) = \begin{cases} x, & \text{if } |x| > c \\ 0, & \text{else} \end{cases}. \qquad (26)$$

Clearly, the domain of this function can be partitioned into three elements, upon two of them the function is the identity function. On the third set, $[-c, c]$, the function is a submersion to a zero-dimensional manifold. We can thus apply Theorem 2 and obtain $l(X \to Y) = P_X([-c, c])$.

It is interesting to see that in both examples the distribution of the input signal does not have an influence on the relative information loss, as long as it is continuous. This is counter-intuitive in the sense that adding two strongly correlated RVs should preserve more information than adding two independent RVs, or, in the sense that center clipping a large signal should not hurt too much. This intuition, however, is based on the fact that one tends to attribute unequal importance to different aspects of the information contained in the input signal (e.g., principle direction, magnitude, etc.).

V. PCA WITH POPULATION COVARIANCE MATRIX

In PCA one uses the eigenvalue decomposition (EVD) of the covariance matrix of a multivariate input to obtain a different representation of the input vector. In particular, let $\mathbf{X}$ be an RV with distribution $P_{\mathbf{X}} \ll \mu^N$ and information dimension $d(\mathbf{X}) = N$. We further assume that $\mathbf{X}$ has zero mean and a positive definite population covariance matrix $\underline{\mathbf{C}}_{\mathbf{X}} = \mathrm{E}\left\{\mathbf{X}\mathbf{X}^T\right\}$ which is known a priori. The case where $\underline{\mathbf{C}}_{\mathbf{X}}$ is not known but has to be estimated from the data is considered in Section VI.

The EVD of the covariance matrix yields

$$\underline{\mathbf{C}}_{\mathbf{X}} = \underline{\mathbf{W}}\underline{\mathbf{\Sigma}}\underline{\mathbf{W}}^T \qquad (27)$$

where $\underline{\mathbf{W}}$ is an orthogonal matrix (i.e., $\underline{\mathbf{W}}^{-1} = \underline{\mathbf{W}}^T$) and $\underline{\mathbf{\Sigma}}$ is a diagonal matrix consisting of the N eigenvalues of $\underline{\mathbf{C}}_{\mathbf{X}}$. We now can describe the PCA by the following linear transform:

$$\mathbf{Y} = \mathbf{g}(\mathbf{X}) = \underline{\mathbf{W}}^T\mathbf{X}. \qquad (28)$$

As in Section IV, the linear transform is bi-Lipschitz, and the information loss vanishes[3].

Often, however, the PCA is used for dimensionality reduction, where after the linear transform in (28) the elements of the random vector $\mathbf{Y}$ with the smallest variances are discarded (thus, preserving the subspace with the largest variance). Essentially, the mapping from $\mathbf{Y}$ to, e.g.,

$$\mathbf{Y}_M = [Y_1, \dots, Y_M] \qquad (29)$$

is a projection onto the first $M < N$ coordinates, which is a submersion between two smooth manifolds. We can thus apply Theorem 2 and obtain the relative information loss $l(\mathbf{Y} \to \mathbf{Y}_M) = \frac{N-M}{N}$. In analogy with the example in Section IV we thus get

$$l(\mathbf{X} \to \mathbf{Y}_M) = \frac{N-M}{N}. \qquad (30)$$

We can now extend this analysis to the case where from $\mathbf{Y}_M$ an (N-dimensional) estimate $\tilde{\mathbf{X}}_M$ of the original data $\mathbf{X}$ is reconstructed. This estimate is obtained using a linear transform

$$\tilde{\mathbf{X}}_M = \underline{\mathbf{W}}\underline{\mathbf{I}}_M^T\mathbf{Y}_M \qquad (31)$$

where $\underline{\mathbf{I}}_M$ is a rectangular identity matrix with M rows and N columns. The (full-rank) matrix $\underline{\mathbf{I}}_M^T$ is a mapping to a higher-dimensional space (i.e., from $\mathbb{R}^M$ to $\mathbb{R}^N$) and is thus bi-Lipschitz; so is the rotation with the matrix $\underline{\mathbf{W}}$. Furthermore, the transform from $\mathbf{Y}_M$ to $\tilde{\mathbf{X}}_M$ is invertible and, as a consequence, no additional information is lost. We thus state

$$l(\mathbf{X} \to \tilde{\mathbf{X}}_M) = \frac{N-M}{N} \qquad (32)$$

where, using above notation, $\tilde{\mathbf{X}}_M = \underline{\mathbf{W}}\underline{\mathbf{I}}_M^T\underline{\mathbf{I}}_M\underline{\mathbf{W}}^T\mathbf{X}$.

Indeed, the same result would have been obtained if the rotation would have been performed using any other orthogonal matrix and regardless which elements of the rotated

---

[3]Not only the relative information loss $l(\mathbf{X} \to \mathbf{Y}) = 0$, but also the absolute information loss $L(\mathbf{X} \to \mathbf{Y}) = 0$.
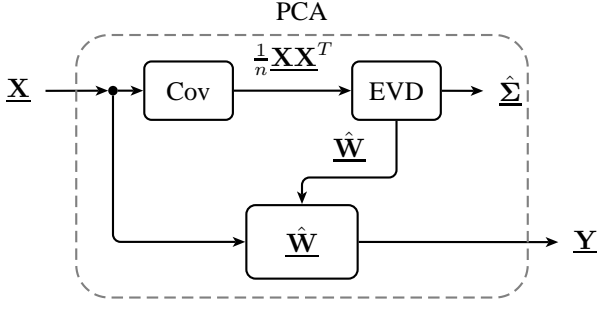
Fig. 2. The PCA as a nonlinear input-output system. "Cov" denotes the computation of the sample covariance matrix and "EVD" stands for eigenvalue decomposition.

vector were discarded. In particular, also if just the first M components of $\mathbf{X}$ would have been preserved, we would have $l(\mathbf{X} \to \mathbf{X}_M) = \frac{N-M}{N}$.

Of course, PCA is known to be optimal in the sense that, by discarding the elements of $\mathbf{Y}$ with the smallest variances, the mean-squared reconstruction error for $\tilde{\mathbf{X}}_M$ is minimized [1]. For this interpretation, measuring the information loss *with respect to a relevant random variable* may do the trick, providing us with a statement about the optimality of PCA in an information-theoretic sense (cf. [1], [4]). Conversely, if one cannot determine which information at the input is relevant, one has no reason to perform PCA prior to reducing the dimension of the data.

VI. PCA WITH SAMPLE COVARIANCE MATRIX

We argued in Section V that the PCA without dimensionality reduction is an invertible transform. This, however, is only true if one has access to the orthogonal matrix $\underline{\mathbf{W}}$; if not, i.e., if one feeds a system with the data matrix $\underline{\mathbf{X}}$ and just receives the rotated matrix $\underline{\mathbf{Y}}$ (see Fig. 2), it can be shown that information is lost. In this section we make this statement precise.

We believe that the following analysis can be generalized to all data matrices with a continuous joint distribution, i.e., $P_{\underline{\mathbf{X}}} \ll \mu^{nN}$. For the sake of simplicity, in this paper, we will focus on a particularly simple scenario: Let $\underline{\mathbf{X}}$ denote a matrix where each of its n columns represents an *independent* sample of an N-dimensional Gaussian RV $\mathbf{X}$. Again, let $\mathbf{X}$ have zero mean and positive definite population covariance matrix $\underline{\mathbf{C}}_\mathbf{X}$. As a consequence, the probability distribution of the data matrix $\underline{\mathbf{X}}$ is absolutely continuous w.r.t. the nN-dimensional Lebesgue measure ($P_{\underline{\mathbf{X}}} \ll \mu^{nN}$).

The sample covariance matrix $\underline{\hat{\mathbf{C}}}_\mathbf{X} = \frac{1}{n}\underline{\mathbf{X}}\underline{\mathbf{X}}^T$ is symmetric and almost surely positive definite. In the usual case where $n \geq N$ one can show that $\frac{N(N+1)}{2}$ entries can be chosen and that the remaining entries depend on these in a deterministic manner. Indeed, since in this case the distribution of $\hat{\mathbf{C}}_\mathbf{X}$ possesses a density (the Wishart distribution, cf. [17]), the distribution is absolutely continuous w.r.t. the Lebesgue measure on an $\frac{N(N+1)}{2}$-dimensional submanifold of the N^2-dimensional Euclidean space. With some abuse of notation we thus write $P_{\underline{\hat{\mathbf{C}}}_\mathbf{x}} \ll \mu^{\frac{N(N+1)}{2}}$.

The orthogonal matrix $\underline{\hat{\mathbf{W}}}$ for PCA (see (28); now applied to the matrix $\underline{\mathbf{X}}$ instead of the vector $\mathbf{X}$) is obtained from the EVD of the sample covariance matrix, i.e.,

$$\underline{\hat{\mathbf{C}}}_\mathbf{X} = \underline{\hat{\mathbf{W}}}\underline{\hat{\boldsymbol{\Sigma}}}\underline{\hat{\mathbf{W}}}^T \tag{33}$$

where $\underline{\hat{\boldsymbol{\Sigma}}}$ is the diagonal matrix containing the eigenvalues of $\underline{\hat{\mathbf{C}}}_\mathbf{X}$. The joint distribution of the N eigenvalues of $\underline{\hat{\mathbf{C}}}_\mathbf{X}$ possesses a density [17, Ch. 9.4]; thus, the distribution of $\underline{\hat{\boldsymbol{\Sigma}}}$ is absolutely continuous w.r.t. the Lebesgue measure on an N-dimensional submanifold of the N^2-dimensional Euclidean space, or $P_{\underline{\hat{\boldsymbol{\Sigma}}}} \ll \mu^N$.

Clearly, the entries of $\underline{\hat{\mathbf{C}}}_\mathbf{X}$ are smooth functions of the eigenvalues and the entries of $\underline{\hat{\mathbf{W}}}$. Images of Lebesgue null-sets under smooth functions between Euclidean spaces of same dimension are null-sets themselves; were the probability measure $P_{\underline{\hat{\mathbf{W}}}}$ supported on some set of dimensionality lower than $\frac{N(N-1)}{2}$, the image of the product of this set and $\mathbb{R}^N$ (for the eigenvalues) would be a Lebesgue null-set with positive probability measure. Since this contradicts the fact that $\underline{\hat{\mathbf{C}}}_\mathbf{X}$ is continuously distributed, it follows that

$$P_{\underline{\hat{\mathbf{W}}}} \ll \mu^{\frac{N(N-1)}{2}}. \tag{34}$$

We now argue (see Appendix for a rigorous discussion) that the rotated data does not tell us anything about the rotation, hence

$$P_{\underline{\hat{\mathbf{W}}}|\underline{\mathbf{Y}}=\underline{\mathbf{y}}} \ll \mu^{\frac{N(N-1)}{2}}. \tag{35}$$

Knowing $\underline{\mathbf{Y}} = \underline{\mathbf{y}}$, $\underline{\mathbf{X}}$ is a linear, bi-Lipschitz function of $\underline{\hat{\mathbf{W}}}$, and thus the information dimension remains unchanged. Therefore we get with Theorem 1

$$l(\underline{\mathbf{X}} \to \underline{\mathbf{Y}}) = \frac{d(\underline{\mathbf{X}}|\underline{\mathbf{Y}})}{d(\underline{\mathbf{X}})} = \frac{N(N-1)}{2nN} = \frac{N-1}{2n}. \tag{36}$$

We now drop a little of the mathematical rigor to analyze, for the sake of completeness, the less common case where there are less data samples than there are dimensions for each sample ($n < N$). In this case, the sample covariance matrix is not full rank, which means that the EVD yields $N - n$ vanishing eigenvalues. Assuming that still $P_{\underline{\hat{\mathbf{C}}}_\mathbf{x}} \ll \mu^{n\frac{2N-n+1}{2}}$ one finds along the same lines as in the case $n \geq N$ that the loss evaluates to

$$l(\underline{\mathbf{X}} \to \underline{\mathbf{Y}}) = \frac{2N-n-1}{2N}. \tag{37}$$

The behavior of the relative information loss as a function of n is shown in Fig. 3 for different choices of N. A worked example for a particularly simple case can be found in the Appendix.

The relative information loss induced by PCA results from the fact that one cannot know *which rotation* led to the output data matrix. As a consequence, the relative information loss decreases with a larger number of samples: the total information increases while the uncertainty about the rotation remains the same. Note further that the relative information loss cannot exceed $\frac{N-1}{N}$ (for $n = 1$), which is due to the fact that the rotation preserves the norm of the sample.
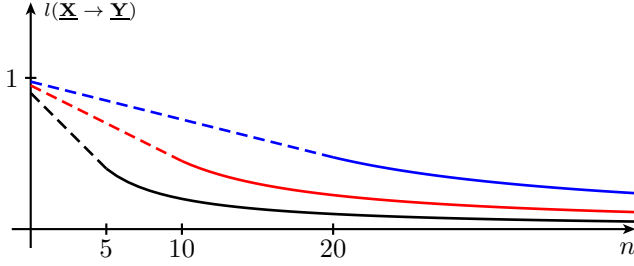
Fig. 3. Relative information loss in the PCA with sample covariance matrix as a function of the number n of independent measurements. The cases $N = 5$ (black), $N = 10$ (red), and $N = 20$ (blue) are shown. The dashed lines indicate the conjectured loss for singular sample covariance matrices.

The PCA with sample covariance matrix also allows different interpretations of information loss, in addition to the information lost in the rotation: First, as we show in the Appendix, by the fact that the sample covariance matrix of $\underline{\mathbf{Y}}$ is a diagonal matrix, the possible values of $\underline{\mathbf{Y}}$ are restricted to a submanifold of dimensionality smaller than nN. Naturally, this also restricts the amount of information which can be conveyed in the output. In contrary to this, in PCA using the population covariance matrix the sample covariance matrix of $\underline{\mathbf{Y}}$ will (almost surely) do not contain zeros, so the entries of $\underline{\mathbf{Y}}$ will not be restricted deterministically.

Finally, it is interesting to observe that in this case the *absolute* information loss in the PCA is infinite (see Proposition 1), even if no additional dimensionality reduction is performed. Moreover, this analysis not only holds for the PCA, but for *any* rotation which depends on the input data in a similar manner – in this sense, the PCA is not better than any other rotation.

VII. DISCUSSION

Relative information loss was introduced to cope with the shortcomings of absolute information loss, especially in cases where the information loss $L(\mathbf{X} \to \mathbf{Y})$ and the information transfer $I(\mathbf{X}; \mathbf{Y})$ are infinite. However, as the examples of Section IV and V showed, even relative information loss not necessarily provides full insight and intuitive interpretations in some cases.

These cases may be characterized by the fact that not all information at the input of the system is *relevant* for a given application, i.e., we are actually not interested in $\mathbf{X}$ itself but in some random variable $\mathbf{Z}$ somehow related to it. $\mathbf{Z}$, however, is not accessible directly, but only through a function $\mathbf{g}$ of the related RV $\mathbf{X}$. The logical consequence is thus to define a quantity which captures the information loss relevant w.r.t. $\mathbf{Z}$. We did so in [18], where we analyzed some of this quantity's properties and made a connection to the signal enhancement problem.

A very similar quantity has already been introduced by Plumbley [4] in the context of unsupervised learning. Using this quantity he proved the optimality of the PCA under appropriate constraints, and argued that information loss in some cases is more versatile than mutual information.

In order to be able to minimize this *relevant information loss*, one clearly has to know something about the relevant

RV $\mathbf{Z}$ and its relationship to the system input $\mathbf{X}$. If one does not have this knowledge, all bits of information have to be treated equally, leading to our notion of (relative) information loss. As a direct consequence, while in some cases the PCA prior to dimensionality reduction might be the optimal solution (cf. [4], [18]), without knowledge about the relevant information one has no reason to apply it to a data set. Even more so, as we showed in Section VI, even without reducing the dimensionality, information can be lost, which should prevent one from unjustified use of PCA.

We now discuss a different aspect of our work about information loss – be it absolute or relative – in deterministic systems. The definition of these quantities allows us to quantify and, hopefully, understand the propagation of information in a network of systems.

Take, for example, the PCA: The information at the input is split and propagates through the system, with parts of it being lost in some paths and preserved in others. By denoting $t(\underline{\mathbf{X}} \to \cdot) = 1 - l(\underline{\mathbf{X}} \to \cdot)$ the relative information transfer, we can thus redraw the system model and obtain Fig. 4 where the arrows are labeled according to the relative information transfer along them. In particular, the eigenvalue decomposition splits the information contained in the covariance matrix into a part describing the eigenvectors and a part describing the eigenvalues. The former part is lost in PCA, while the second is preserved in the output. However, as we already mentioned before, the rotation (i.e., the multiplication with the orthogonal matrix) removes exactly as much information from the input as is contained in the orthogonal matrix. The information contained in both $\underline{\mathbf{Y}}$ and $\hat{\underline{\mathbf{W}}}$ suffices to reconstruct $\underline{\mathbf{X}}$. If we perform in addition a sphering transform on $\underline{\mathbf{Y}}$ (thus making all eigenvalues unity), one needs the triple $\tilde{\underline{\mathbf{Y}}}$, $\hat{\underline{\mathbf{W}}}$ and $\hat{\underline{\mathbf{\Sigma}}}$ to reconstruct $\underline{\mathbf{X}}$.

It is obvious that this transfer graph must not be understood in the sense of a preservation theorem, similar to Kirchhoff's current law: Information can be split and fused, and after splitting the sum of information at the output needs not equal the sum of information at the input (as it is, by coincidence, for the EVD). If the output information is less than the input information, the intermediate system destroyed the remaining information (as in rotation). Conversely, the if the sum of output information exceeds the input information, this only means that parts of these information must be the same (as, e.g., $\underline{\mathbf{Y}}$ includes the knowledge about $\hat{\underline{\mathbf{\Sigma}}}$).

VIII. CONCLUSION

We have introduced the notion of relative information loss to analyze systems for which the output has a lower dimension than the input, exploiting a tight connection to Rényi's information dimension. As a first result, we showed that the relative information loss for dimensionality reduction is not affected by performing a principle component analysis (PCA) beforehand.

We then showed that even without dimensionality reduction the PCA is not information lossless, given that the sample covariance matrix is used to compute the rotation matrix.
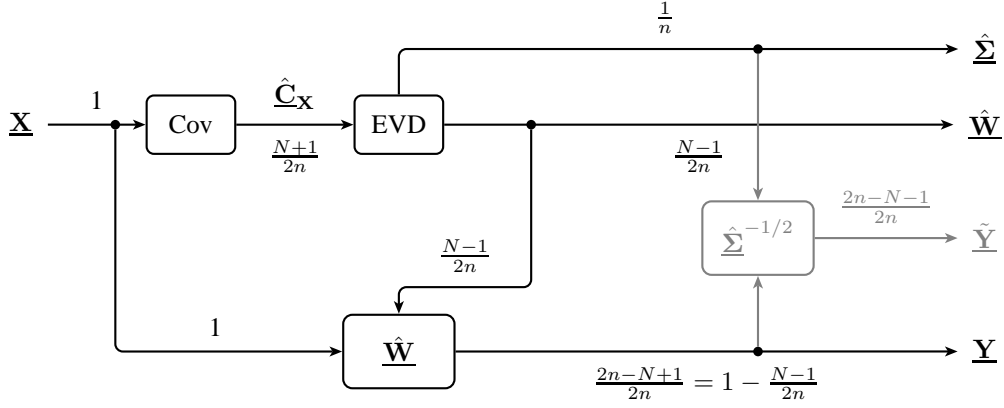
Fig. 4. Information propagation in PCA with sample covariance matrix. The values on the arrows indicate the relative information transfer $t(\underline{\mathbf{X}} \to \cdot)$. We assume $n > N$ in this case. The gray part considers the sphering transform (see text). Note that the separate information transfers to $\underline{\hat{\boldsymbol{\Sigma}}}$, $\underline{\hat{\mathbf{W}}}$, and $\underline{\mathbf{Y}}$ add up to $1 + \frac{1}{n} > 1$. This is because the information in $\underline{\hat{\boldsymbol{\Sigma}}}$ is already contained in $\underline{\mathbf{Y}}$. Note further that $t(\underline{\mathbf{X}} \to \underline{\mathbf{Y}}, \underline{\hat{\boldsymbol{\Sigma}}}, \underline{\hat{\mathbf{W}}}) = 1$, as expected.

There, the relative information loss appears to decrease with increasing sample size.

We proposed to use these somewhat counter-intuitive results as motivation to introduce a notion of information loss which takes the relevance of the data into account, and thus connects to results available in the literature. A detailed analysis of this relevant information loss is within the scope of future work.

APPENDIX A
EXAMPLE FOR INFINITE ABSOLUTE AND VANISHING RELATIVE INFORMATION LOSS

We now give an example where an infinite amount of information is lost (i.e., $L(X \to Y) = \infty$), but for which the relative information loss nevertheless vanishes (i.e., $l(X \to Y) = 0$). To this end, assume that we consider the following scalar function $g \colon (0, 1] \to (0, 1]$

$$g(x) = 2^n(x - 2^{-n}) \text{ if } x \in (2^{-n}, 2^{-n+1}], \ n \in \mathbb{N} \quad (38)$$

In other words, this function maps every interval $(2^{-n}, 2^{-n+1}]$ onto the interval $(0, 1]$. Assume now further that the input variable X has a continuous distribution with density function

$$f_X(x) = 2^n \left(\frac{1}{\log(n+1)} - \frac{1}{\log(n+2)} \right),$$
$$\text{if } x \in (2^{-n}, 2^{-n+1}], \ n \in \mathbb{N} \quad (39)$$

where $\log$ denotes the binary logarithm. As an immediate consequence, the output RV Y is uniformly distributed on $(0, 1]$. Furthermore, the information dimension of the input is given by $d(X) = 1$.

Since the function is piecewise strictly monotone we can apply the reasoning of [5] and claim that

$$L(X \to Y) = H(X|Y) = H(W|Y) \quad (40)$$

where

$$W = n \text{ if } X \in (2^{-n}, 2^{-n+1}]. \quad (41)$$

In particular, for the given input density function we obtain

$$\Pr(W = n|Y = y) = \Pr(W = n)$$
$$= \frac{1}{\log(n+1)} - \frac{1}{\log(n+2)}. \quad (42)$$

For this distribution, however, it is known that the entropy is infinite, i.e., $H(W) = \infty$ [19]. This shows that in this case the absolute information loss is infinite as well.

However, for every $y \in (0, 1]$ the preimage is a countable set, thus $X|Y = y$ is a discrete RV. In addition to that, since X is supported on a compact set, so is $X|Y = y$, and every quantization $\hat{X}_n|Y = y$ will have finite entropy. Thus, $d(X|Y = y) = 0$ for all y [8], and with Theorem 1 we obtain $l(X \to Y) = 0$.

APPENDIX B
PROOF OF CONJECTURE 1 FOR DISCRETE RVS

Let $\mathbf{X}$, $\mathbf{Y}$, and $\mathbf{Z}$ be discrete RVs with finite entropy, and let further $\mathbf{Y} = \mathbf{g}(\mathbf{X})$ and $\mathbf{Z} = \mathbf{h}(\mathbf{Y})$. As an immediate result, $I(\mathbf{X}; \mathbf{Y}) = H(\mathbf{Y})$ and $I(\mathbf{Y}; \mathbf{Z}) = H(\mathbf{Z})$. We next note that

$$l(\mathbf{X} \to \mathbf{Z}) = \frac{H(\mathbf{X}|\mathbf{Z})}{H(\mathbf{X})} \quad (43)$$

since the RVs are already discrete. Furthermore, we can write

$$l(\mathbf{X} \to \mathbf{Z}) = \frac{H(\mathbf{X}) - I(\mathbf{X}; \mathbf{Z})}{H(\mathbf{X})} = 1 - t(\mathbf{X} \to \mathbf{Z}) \quad (44)$$

where $t(\mathbf{X} \to \mathbf{Z}) = \frac{I(\mathbf{X}; \mathbf{Z})}{H(\mathbf{X})} = \frac{H(\mathbf{Z})}{H(\mathbf{X})}$, since also $\mathbf{Z}$ is a function of $\mathbf{X}$. Expanding the latter term – the *relative information*

transfer – one obtains

$$1 - l(\mathbf{X} \to \mathbf{Z}) = t(\mathbf{X} \to \mathbf{Z}) = \frac{H(\mathbf{Z})}{H(\mathbf{X})} \frac{H(\mathbf{Y})}{H(\mathbf{Y})}$$
$$= \frac{I(\mathbf{Y};\mathbf{Z})}{H(\mathbf{Y})} \frac{I(\mathbf{X};\mathbf{Y})}{H(\mathbf{X})} = t(\mathbf{X} \to \mathbf{Y})t(\mathbf{Y} \to \mathbf{Z})$$
$$= (1 - l(\mathbf{X} \to \mathbf{Y}))(1 - l(\mathbf{Y} \to \mathbf{Z}))$$
$$= 1 - l(\mathbf{X} \to \mathbf{Y}) - l(\mathbf{Y} \to \mathbf{Z}) + l(\mathbf{X} \to \mathbf{Y})l(\mathbf{Y} \to \mathbf{Z}).$$
$$(45)$$

Rearranging completes the proof. ∎

We note in passing that this result holds for the cascade of *deterministic* systems and that a generalization to Markov chains $\mathbf{X} - \mathbf{Y} - \mathbf{Z}$ is not possible.

APPENDIX C
INFORMATION DIMENSION OF $\underline{\mathbf{X}}|\underline{\mathbf{Y}} = \underline{\mathbf{y}}$

We already observed that $P_{\hat{\underline{\mathbf{W}}}} \ll \mu^{\frac{N(N-1)}{2}}$. Furthermore, since we know that the sample covariance matrix of $\underline{\mathbf{Y}}$ is a diagonal matrix, the corresponding equations

$$\forall 1 \le i < j \le N: \quad (\underline{\mathbf{C}}_{\mathbf{Y}})_{ij} = \frac{1}{n} \sum_{k=1}^{n} Y_{ik} Y_{jk} = 0 \quad (46)$$

restrict the possible values of $\underline{\mathbf{Y}}$ from an nN-dimensional to an M-dimensional subspace with

$$M = nN - \frac{N(N-1)}{2}. \quad (47)$$

In fact, it can be shown that M elements of $\underline{\mathbf{Y}}$ are random while the remaining $(nN - M)$ depend on these in a deterministic manner[4].

Since $\underline{\mathbf{X}} = \hat{\underline{\mathbf{W}}}\underline{\mathbf{Y}}$, $\underline{\mathbf{X}}$ is a smooth function from $\mathbb{R}^{\frac{N(N-1)}{2}} \times \mathbb{R}^M$ (the ranges of the values of $\hat{\underline{\mathbf{W}}}$ and $\underline{\mathbf{Y}}$) to $\mathbb{R}^{nN}$ (the range of values of $\underline{\mathbf{X}}$), thus from $\mathbb{R}^{nN}$ to $\mathbb{R}^{nN}$. Since this smooth mapping maps null-sets to null-sets [13, Lem. 6.2], we obtain

$$P_{(\hat{\underline{\mathbf{W}}}, \underline{\mathbf{Y}})} \ll \mu^{nN}. \quad (48)$$

(We are well aware that $\hat{\underline{\mathbf{W}}}$ and $\underline{\mathbf{Y}}$ together have more than nN entries, but only nN of those can be chosen freely. In other words, the graph of the functions defining the remaining entries of $\underline{\mathbf{Y}}$ and $\hat{\underline{\mathbf{W}}}$ is an nN-dimensional submanifold of $\mathbb{R}^{N(n+N)}$ [13, Lem. 5.9].) The joint distribution of $(\hat{\underline{\mathbf{W}}}, \underline{\mathbf{Y}})$ thus possesses a density, and by marginalizing and conditioning so does $\hat{\underline{\mathbf{W}}}|\underline{\mathbf{Y}} = \underline{\mathbf{y}}$. As a consequence,

$$P_{\hat{\underline{\mathbf{W}}}|\underline{\mathbf{Y}}=\underline{\mathbf{y}}} \ll \mu^{\frac{N(N-1)}{2}}. \quad (49)$$

Note further that this does not mean that $\hat{\underline{\mathbf{W}}}$ is independent of $\underline{\mathbf{Y}}$ – it just means that these two quantities are at least not related deterministically.

The final step is taken by recognizing that if one knows $\underline{\mathbf{Y}} = \underline{\mathbf{y}}$, then $\underline{\mathbf{X}}|\underline{\mathbf{Y}} = \underline{\mathbf{y}}$ is a linear function of $\hat{\underline{\mathbf{W}}}|\underline{\mathbf{Y}} = \underline{\mathbf{y}}$. Since $\underline{\mathbf{Y}}$ has full rank, the linear function maps the N^2-dimensional space of $\hat{\underline{\mathbf{W}}}$ (on which the probability mass is concentrated

---
[4]E.g., one could determine Y_{ij} from the equation for $(\underline{\mathbf{C}}_{\mathbf{Y}})_{ij}$.

on an $\frac{N(N-1)}{2}$-dimensional subspace) to the N^2-dimensional linear subspace of $\mathbb{R}^{nN}$. With [20, Remark 28.9] this transform is bi-Lipschitz and preserves the information dimension. Thus,

$$d(\underline{\mathbf{X}}|\underline{\mathbf{Y}} = \underline{\mathbf{y}}) = \frac{N(N-1)}{2}. \quad (50)$$

APPENDIX D
EXAMPLE: PCA WITH SINGULAR SAMPLE COVARIANCE MATRIX

We now give a worked example for the – admittedly less common – case of a singular sample covariance matrix. Let $\mathbf{X}$ be a two-dimensional Gaussian RV, and let $n = 1$, i.e., $\underline{\mathbf{X}} = \mathbf{X}$. The sample covariance matrix is given by

$$\hat{\underline{\mathbf{C}}}_{\mathbf{X}} = \begin{bmatrix} X_1^2 & X_1 X_2 \\ X_1 X_2 & X_2^2 \end{bmatrix} \quad (51)$$

and has eigenvalues $|\mathbf{X}|^2$ and 0. The corresponding (normalized) eigenvectors are then given by

$$\mathbf{p}_1 = \left[\frac{\operatorname{sgn}(X_2) X_1}{|\mathbf{X}|}, \frac{|X_2|}{|\mathbf{X}|} \right]^T \quad (52)$$

and

$$\mathbf{p}_2 = \left[-\frac{\operatorname{sgn}(X_1) X_2}{|\mathbf{X}|}, \frac{|X_1|}{|\mathbf{X}|} \right]^T. \quad (53)$$

Performing now the rotation $\mathbf{Y} = \hat{\underline{\mathbf{W}}}^T \mathbf{X}$ with $\hat{\underline{\mathbf{W}}} = [\mathbf{p}_1, \mathbf{p}_2]$ one obtains

$$\mathbf{Y} = [\operatorname{sgn}(X_2) |\mathbf{X}|, 0]^T. \quad (54)$$

The fact that the second component of $\mathbf{Y}$ is zero regardless of the entries of $\mathbf{X}$ makes it obvious that exactly one half of the information is lost, i.e., $l(\mathbf{X} \to \mathbf{Y}) = \frac{1}{2}$. This also corresponds to the result obtained in Section VI for $N = 2$ and $n = 1$.

REFERENCES

[1] G. Deco and D. Obradovic, *An Information-Theoretic Approach to Neural Computing*. New York, NY: Springer, 1996.

[2] J. C. Principe, *Information Theoretic Learning: Renyi's Entropy and Kernel Perspectives*, ser. Information Science and Statistics. New York, NY: Springer, 2010.

[3] R. Linsker, "Self-organization in a perceptual network," *IEEE Computer*, vol. 21, no. 3, pp. 105–117, Mar. 1988.

[4] M. Plumbley, "Information theory and unsupervised neural networks," Cambridge University Engineering Department, Tech. Rep. CUED/F-INFENG/TR. 78, 1991.

[5] B. C. Geiger and G. Kubin, "On the information loss in memoryless systems: The multivariate case," in *Proc. Int. Zurich Seminar on Communications (IZS)*, Zurich, Feb. 2012, pp. 32–35, extended version available: arXiv:1109.4856 [cs.IT].

[6] G. H. Golub and C. F. van Loan, *Matrix Computations*, 3rd ed. Baltimore, MD: John Hopkins University Press, 1996.

[7] M. S. Pinsker, *Information and Information Stability of Random Variables and Processes*. San Francisco, CA: Holden Day, 1964.

[8] A. Rényi, "On the dimension and entropy of probability distributions," *Acta Mathematica Hungarica*, vol. 10, no. 1-2, pp. 193–215, Mar. 1959.

[9] A. N. Kolmogorov, "On the Shannon theory of information transmission in case of continuous signals," *IEEE Trans. Inf. Theory*, vol. 2, no. 4, pp. 102–108, Dec. 1956.

[10] ——, "ε-entropy and ε-capacity of sets in functional spaces," in *Selected Works of A. N. Kolmogorov – Volume III: Information Theory and the Theory of Algorithms*, A. N. Shiryayev, Ed. Dordrecht: Kluwer, 1993, pp. 86–170.

[11] W. Rudin, *Real and Complex Analysis*, 3rd ed. New York, NY: McGraw-Hill, 1987.

[12] Y. Wu and S. Verdú and, "Rényi information dimension: Fundamental limits of almost lossless analog compression," *IEEE Trans. Inf. Theory*, vol. 56, no. 8, pp. 3721–3748, Aug. 2010.

[13] J. M. Lee, *Introduction to Smooth Manifolds*, ser. Graduate Texts in Mathematics. New York, NY: Springer, 2003.

[14] S. P. Ponomarev, "Submersions and preimages of sets of measure zero," *Siberian Mathematical Journal*, vol. 28, no. 1, pp. 153–163, Jan. 1987.

[15] M. Śmieja and J. Tabor, "Entropy of the mixture of sources and entropy dimension," *IEEE Trans. Inf. Theory*, vol. 58, no. 5, pp. 2719–2728, May 2012.

[16] P. Vary and R. Martin, *Digital speech transmission: Enhancement, coding and error concealment*. Chichester: John Wiley & Sons, 2006.

[17] R. J. Muirhead, *Aspects of Multivariate Statistical Theory*, ser. Wiley Series in Probability and Mathematical Statistics. Hoboken, NJ: Wiley Interscience, 1982.

[18] B. C. Geiger and G. Kubin, "Signal enhancement as minimization of relevant information loss," May 2012, `arXiv:1205.6935 [cs.IT]`.

[19] M. Baer, "A simple countable infinite-entropy distribution," 2008. [Online]. Available: https://hkn.eecs.berkeley.edu/~calbear/research/Hinf.pdf

[20] J. Yeh, *Lectures on real analysis*. Singapore: World Scientific Publishing, 2000.